Análise Discriminante Múltipla

TUANY DE PAULA CASTRO

Finalidades e Aplicações

Discriminação e classificação são técnicas que concernem em separar grupos de objetos distintos e alocar novos objetos em grupos previamente definidos.

- ✓ **Objetivo 1:** Descrever graficamente ou algebricamente as características distintas dos grupos. Encontrar os “discriminantes”.
- ✓ **Objetivo 2:** Derivar uma regra de classificação para novos objetos.

Quando usar?
Quando temos interesse em prever a que grupo uma determinada observação (empresa, pessoa, cliente, produto) pertence.

Finalidades e Aplicações

Diferença entre Análise Discriminante, Regressão Linear Múltipla e Regressão Logística

Técnica	Variável Dependente	Variáveis Independentes
Regressão Linear Múltipla	Quantitativa	Quantitativas ou qualitativas
Regressão Logística	Qualitativa (duas categorias)	Quantitativas ou qualitativas
Análise Discriminante	Qualitativa	Quantitativas

Finalidades e Aplicações

Análise Discriminante

Simples

- Variável dependente com 2 categorias

Múltipla

- Variável dependente com 3 ou mais categorias

Finalidades e Aplicações

Populações	Variáveis independentes
Pessoas com problemas de estômago e pessoas normais	Medidas de ansiedade, culpa, perfeccionismo
Textos dos autores Machado de Assis, Paulo Coelho e Fernando Pessoa	Frequências de diferentes palavras e tamanhos das sentenças
Bons e maus riscos de crédito	Renda, idade, número de cartões de crédito, tamanho da família

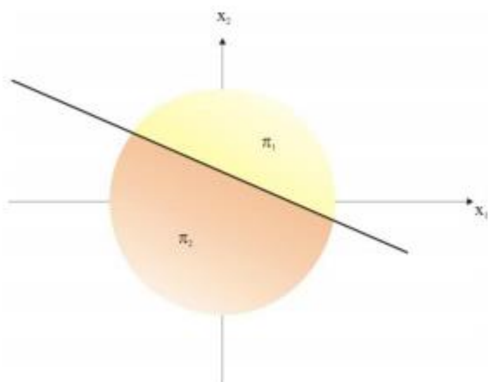
Finalidades e Aplicações

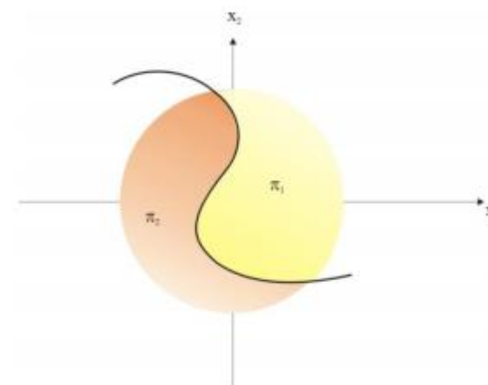
O problema consiste em obter uma combinação linear de características observadas que apresente maior poder de discriminação entre populações.

- Esta combinação linear é a função discriminante.

- Esta função tem a propriedade de minimizar as probabilidades de má classificação.

- Ela permite a obtenção de fronteiras para classificação.
- Essas fronteiras não são exatamente definidas, havendo superposição, e consequentemente, erro de classificação.

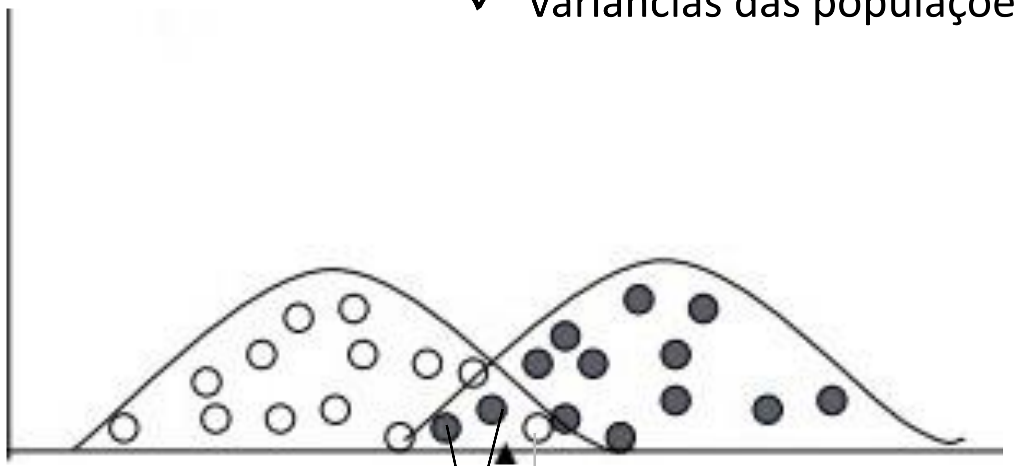




Finalidades e Aplicações

A Análise Discriminante será mais assertiva quando:

- ✓ Médias das populações forem distantes
- ✓ Variâncias das populações forem pequenas



Erro: observações do grupo 2, mas classificadas em 1

Erro: observação do grupo 1, mas classificada em 2

História

Em 1930, três diferentes pessoas (R.A. Fisher no Reino Unido, Hotelling nos EUA e Mahalanob na Índia) estavam tentando resolver o mesmo problema por três abordagens diferentes. Mais tarde, seus métodos (Função Discriminante Linear de Fisher, teste T^2 de Hotelling e distância de Mahalanobis) foram combinadas para formar o que chamamos hoje de Análise Discriminante.

Exemplo Motivador

Uma grande transportadora aérea internacional recolheu dados sobre os empregados em três classificações distintas:

- Serviço ao cliente (*customer service)*
- Mecânica (*mechanics)*
- Despachante (*dispatchers)*

Como diferir os grupos?

Análise Discriminante!

O diretor de Recursos Humanos quer saber se estas três classificações de trabalho estão ligadas a diferentes tipos de personalidade. Com esse objetivo, a cada funcionário foi administrada uma bateria de testes psicológicos que incluem medidas de interesse na atividade ao ar livre, sociabilidade e conservatividade.

Suposições

- ✓ **Tamanho adequado da amostra:** pelo menos 20 casos (n) para cada grupo (número de observações maior do que número de variáveis independentes).

- ✓ **Multicolinearidade:** variáveis independentes com valores altos de correlação.

- ✓ **Método Geral:** suposição de distribuição Normal multivariada para as variáveis independentes.

- ✓ **Método de Fisher:** suposição de homogeneidade de variâncias/covariâncias.

Verificação das suposições

✓ **Método Geral:** suposição de distribuição Normal multivariada para as variáveis independentes.

- Teste de Normalidade

$$H_0: os\ dados\ seguem\ uma\ distribuição\ normal\ multivariada$$
$$H_1: os\ dados\ não\ seguem\ uma\ distribuição\ normal\ multivariada$$

Verificação das suposições

✓ **Método de Fisher:** suposição de homogeneidade de variâncias/covariâncias.

- Teste de Box

$$H_0: matrizes\ de\ variâncias\ e\ covariâncias\ iguais$$
$$H_1: pelo\ menos\ uma\ matriz\ diferente$$

Exemplo - Suposições

- Tamanho da amostra

```
Console ~/
> # Verificando tamanho dos grupos
> table(dados$job)

customer service         dispatch         mechanic
              85               66               93
>
```

n > 20 para todos os grupos

- Método de Fisher: homogeneidade das variâncias

```
Console F:/
> boxM(data = dados, grouping = dataset$job)

        Box's M-test for Homogeneity of Covariance Matrices

data:   dados
Chi-Sq (approx.) = 25.642, df = 12, p-value = 0.01206
```

Valor-p > 1%, então variâncias iguais

Exemplo - Suposições

- Multicolinearidade

```
Console ~/
> # Matriz de correlações (suposição de não multicolinearidade)
> cor(dados[,1:3])
                 outdoor      social conservative
outdoor       1.00000000 -0.07130338   0.07938108
social       -0.07130338  1.00000000  -0.23586453
conservative  0.07938108 -0.23586453   1.00000000
> 
```

Correlações baixas

Métodos de seleção das variáveis

De que maneira determinar as variáveis que entrarão na análise?

- ✓ **Método Enter**
  - Incorpora todas as variáveis no modelo
  - Deve ser utilizado quando se tem certeza de que todas as variáveis são necessárias para o estudo

- ✓ **Método Stepwise:**
  - Incorpora no modelo, passo a passo, variável independente que mais distingue os grupos.

Métodos de seleção das variáveis

✓ **Método Stepwise:** acaba quando uma das quatro condições abaixo ocorre:

1. Todas as variáveis foram incluídas ou removidas.
2. O número máximo de passos foi atingido.
3. Nenhuma outra variável que não está no modelo tem o valor de F maior do que o valor F de entrada.
4. Qualquer variável após o passo tem um valor de tolerância menor do que o especificado.

- **Forward Stepwise:** inclusão da variável com maior valor F (maior do que o F especificado para inclusão).

- **Backward Stepwise:** exclusão da variável com menor valor F (menor do que o F especificado para exclusão).

Métodos de seleção das variáveis

- **Lambda de Wilks:** estatística para avaliar o poder discriminatório de cada variável no modelo:
  - Próximo de 1: nenhum poder discriminatório
  - Próximo de 0: perfeito poder discriminatório

- **Valor F e valor-p:** o valor Lambda de Wilks pode ser convertido no valor F para se obter o valor-p. Se valor-p < 0.01, então variável é incluída no modelo.

- **R-quadrado:** indicação da associação da variável com a resposta.

Exemplo – seleção de variáveis

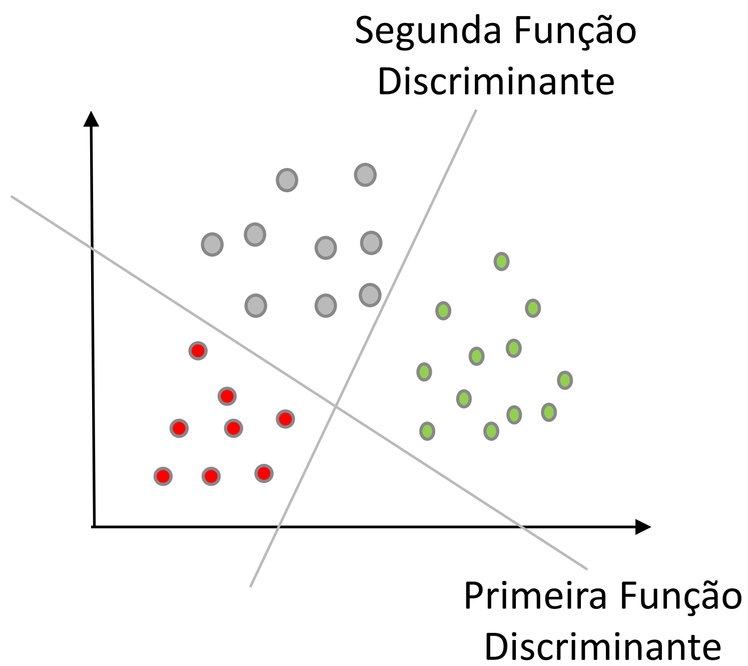
```
> library(DiscriMiner)
> ########### Seleção das variáveis
> # Medidas de qualidade
> discPower(variables = dados[,1:3], group = dados$job)
             correl_ratio wilks_lambda F_statistic      p_value
outdoor         0.2828048    0.7171952    47.51563 0.000000e+00
social          0.3960186    0.6039814    79.00945 0.000000e+00
conservative    0.2049648    0.7950352    31.06561 9.920953e-13
>
```

Observamos que a variável Sociabilidade é a que mais contribui na discriminação dos grupos (menor Lambda de Wilks e maior R^2), seguida pela Atividades ao ar livre e por último Conservatividade.

Como os valores-p dos testes de significância, foram todos menores do que 1%, vamos incluir todas as variáveis no modelo.

Função Discriminante - Método de Fisher

A motivação de Fisher foi a necessidade de obter uma representação razoável das populações envolvendo apenas algumas combinações lineares das observações.



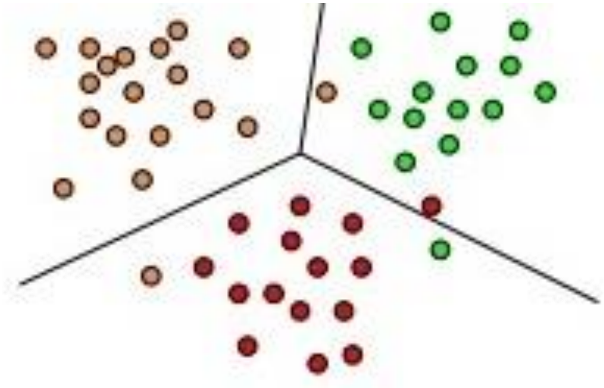
- As **funções discriminantes** são independentes
- O número máximo de funções discriminantes é igual a:
  Mínimo(g-1, p)

Com g o número de grupos e p o número de variáveis independentes

Função Discriminante - Método de Fisher

As **funções discriminantes** são obtidas garantindo:

- distância máxima entre as médias dos grupos
- variância mínima dentro dos grupos



Função Discriminante - Método de Fisher

Uma **função discriminante** é uma função linear do tipo:

$$Y = a + \ b_1 * x_1 + b_2 * x_2 + \cdots + b_m * x_m$$

- a uma constante
- x_1 a x_m as variáveis independentes
- b_1 a b_m os coeficientes da discriminante de cada variável independente. Seus valores indicam o grau de contribuição para a predição do grupo.

Exemplo – Funções Discriminantes

Funções discriminantes determinadas para o exemplo:

```
Console ~/
> discrim_1 <- desDA(variables = dados[, c("outdoor", "social", "conservative")],
+                    group = dados$job)
> # Coeficientes das funções discriminantes
> discrim_1$discrivar
                    DF1         DF2
constant     -1.29771533 -3.58212448
outdoor      -0.06881073  0.22858100
social        0.19830515  0.04181027
conservative -0.16301284 -0.08094410
> 
```

$$DF_1 = -1{,}30 - 0{,}07 * Outdoor + 0{,}20 * Social - 0{,}16 * Conservative$$

$$DF_2 = -3{,}58 + 0{,}23 * Outdoor + 0{,}04 * Social - 0{,}08 * Conservative$$

Exemplo – Funções Discriminantes

Os autovalores indicam quão distintos os grupos são nas funções discriminantes, ou seja, quanto mais afastados de 1, maiores serão as variações entre os grupos explicadas pelas funções discriminantes.

A primeira função discriminante explica 79% da variabilidade dos dados.

```
Console ~/
> # Autovalores das funções discriminantes e variabilidade explicada
> discrim_1$values
        value proportion accumulated
DF1 1.3431063   79.52844    79.52844
DF2 0.3457314   20.47156   100.00000
>
```

Função Discriminante - Método de Fisher

A **matriz de Fatores** representa as correlações entre as variáveis e as funções discriminantes e são usadas para interpretação das funções.

Exemplo:

```
Console ~/
> # Matriz de fatores para interpretação
> discrim_1$discor
                   DF1        DF2
outdoor      -0.3199511  0.9310830
social        0.8743335  0.1933344
conservative -0.6208989 -0.2339640
```

A primeira função discriminante é mais fortemente ponderada pela Sociabilidade (*Social)* e em menor medida por Conservatividade (*Conservative)* no sentido contrário. Já a segunda função discriminante é mais marcada pela Atividade ao ar livre (*Outdoor).*

Função Discriminante - Método de Fisher

O **centróide** de cada grupo é a média das funções discriminantes em cada grupo. Obtido quando utilizamos a média de cada variável independente dentro de cada grupo g para cada função discriminante.

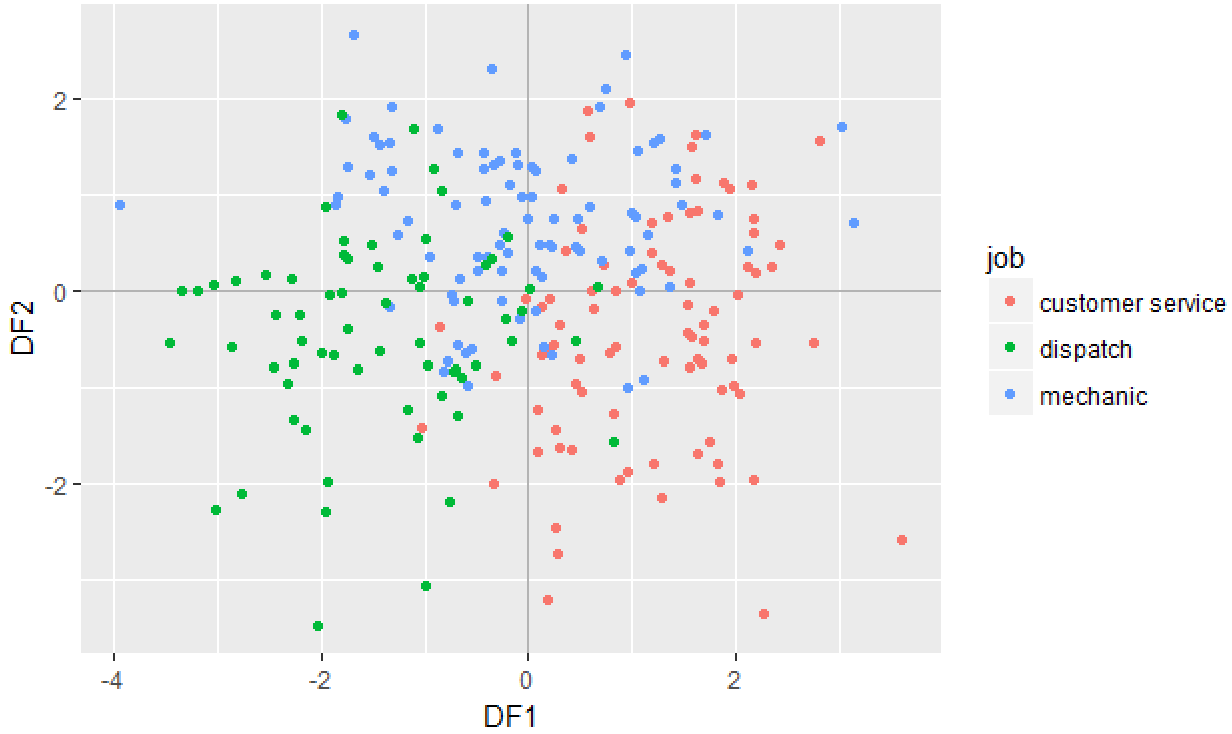
Exemplo:

```
Console ~/
> # Centróides
> dados$escores1 <- discrim_1$scores[,1]
> dados$escores2 <- discrim_1$scores[,2]
> dados %>% group_by(job) %>% summarise(C1 = mean(escores1), C2 = mean(escores2))
# A tibble: 3 x 3
               job          C1         C2
            <fctr>       <dbl>      <dbl>
1 customer service  1.17363564 -0.4384373
2         dispatch -1.46346386 -0.4475396
3         mechanic -0.03409048  0.7183310
>
```

Aparentemente, a primeira função discriminante discrimina mais entre Despachante e outros grupos, pois a média de Despachante é razoavelmente diferente das demais. Já a segunda função discriminante parece distinguir mais entre Mecânica e outros.

Função Discriminante - Método de Fisher

Uma maneira rápida de ver a análise dos centroides é construindo um scatterplot para as duas funções discriminantes.



Exercícios

1) Admita dois grupo de objetos (A e B) para os quais obtivemos os valores de duas variáveis independentes ($X_1 e\ X_2$). Obtenha e represente a função discriminante.

$$Grupo\ A \begin{cases} X_1: 3, 2, 4 \\ X_2: 7, 4, 7 \end{cases} \text{ e } Grupo\ B \begin{cases} X_1: 6, 2, 4 \\ X_2: 9, 7, 8 \end{cases}$$

2) Considere a observação de duas variáveis independentes de 3 populações. Assuma que as populações têm matrizes de covariâncias iguais. Represente as funções discriminantes em um gráfico.

$$População\ 1 \begin{cases} X_1: -2, 0, -1 \\ X_2: 5, 3, 1 \end{cases},\ População\ 2 \begin{cases} X_1: 0, 2, 1 \\ X_2: 6\ 4, 2 \end{cases}\ ePopulação\ 3 \begin{cases} X_1: 1, 0, -1 \\ X_2: -2, 0, -4 \end{cases}$$

Testes de Significância das Funções Discriminantes

O **teste Lambda de Wilks** testa globalmente o poder discriminatório das funções discriminantes.

$H_0: As\ funções\ discriminantes\ não\ são\ significantes\ para\ discriminar\ os\ grupos$

$H_1: As\ funções\ discriminantes\ são\ significantes\ para\ discriminar\ os\ grupos$

Queremos rejeitar H_0. Quanto mais próximo de zero o valor do Lambda de Wilks, mais discriminatórias são as funções discriminantes.

Exemplo – Teste de significância

Console ~/

```
> summary(manova(escores ~ Y_real), test="Wilks")
           Df   Wilks approx F num Df den Df    Pr(>F)    
Y_real      2 0.36399   78.901      4    480 < 2.2e-16 ***
Residuals 241                                              
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pelo Lambda de Wilks, podemos concluir que ambas funções discriminantes são significantes para o estudo. Além disso, o valor-p < 0,01 nos leva a rejeição de H_0 e concluímos que ambas funções discriminantes são significativas.

Exercícios

3) Considere os dados referentes a duas raças de insetos (A e B) com variáveis independentes o número médio de cerdas primordiais (CP) e o número médio de cerdas distais (CD). Os dados se encontram na planilha *dados_insetos.xlsx*.

- (a) Antes de inicializar a análise, quantas funções discriminantes serão obtidas?
- (b) Determine e interprete as funções discriminantes dessas duas raças pelo método de Fisher selecionando todas as variáveis para a análise.
- (c) Analise o poder de discriminação das variáveis.
- (d) Analise a significâncias das funções discriminantes.

Exercícios

4) A planilha *dados_cereais.xlsx* contém dados sobre cereais matinais produzidos por três diferentes fabricantes americanos: General Mills (G), Kellogg (k) e Quaker (Q). Assuma que os dados seguem Normal Multivariada com uma matriz de covariância comum. Queremos discriminar os três fabricantes com as variáveis Calorias, Proteína, Gordura, Sódio, Fibra, Carboidratos, Açúcar e Potássio. Responda:

- (a) Quais variáveis discriminam melhor os grupos?
- (b) Antes de realizar a análise, quantas funções discriminantes você obterá?
- (c) Determine e interprete as funções discriminantes. Avalie também a significância das funções e comente.
- (d) Parece que algum fabricante está associado com mais “cereais nutritivos” (maiores quantidades de proteína e fibra, menores quantidades de gordura e açúcar e assim por diante) do que outros?
- (e) Construa o gráfico dos cereais no espaço bidimensional discriminante usando diferentes símbolos para identificar os três fabricantes. Interprete.

Classificação

A partir da observação de um conjunto de variáveis, é possível identificar a qual grupo uma unidade amostral deve pertencer?

Sim!
A partir da Análise Discriminante, pode-se discriminar grupos e prever a classificação de novas observações.

Classificação

Há diferentes critérios de classificação:

- Classificação *a priori*
- Classificação Post – hoc
- Funções de Classificação
- Distância de Mahalanobis
- Classificação de casos

Classificação

- **Classificação *a priori***
  - Probabilidades de um caso pertencer a um grupo sem qualquer conhecimento das variáveis do modelo.
  - Exemplo: você pode saber *a priori* que há mais pessoas de Serviço ao Cliente na empresa e, por isso, a probabilidade *a priori* de uma pessoa pertencer a esse grupo é maior do que para qualquer outro grupo.
  - Essas probabilidades podem afetar muito a precisão da classificação.

Classificação

- **Classificação Post – hoc**
  - Classificações obtidas post-hoc, ou seja, após análise e não predições a priori (post-hoc = depois disso)
  - Essas classificações devem ser vistas como ferramentas de diagnóstico (identificar pontos fortes e fracos nas funções de classificação)
  - Espera-se menos precisão na classificação de novos casos.

Classificação

- **Funções de Classificação**
  - Utilizadas para determinar a que grupo cada caso pertence.
  - Há tantas funções de classificação quanto grupos.
  - Cada função permite calcular um escore de classificação para cada caso em cada grupo. Classifica-se o caso no grupo em que apresentou o escore mais elevado.

Exemplo - Classificação

Funções de classificação:

```
Console ~/
> fit <- linDA(variables = dados[, c("outdoor", "social", "conservative")],
+              group = dados$job)
>
> # Funções de classificação
> fit$functions
             customer service    dispatch    mechanic
constant          -23.0975206 -20.7714819 -25.2062526
outdoor             0.6262915   0.8427028   0.9965617
social              1.2406066   0.7221426   1.0379702
conservative        0.6864765   1.1056750   0.7955794
> 
```

Em qual grupo você classificaria um funcionário que apresentou 22 de Sociabilidade, 10 de Atividades ao ar livre e 5 de Conservatividade?

Exemplo – Classificação

```
Console ~/
> fit <- linDA(variables = dados[, c("outdoor", "social", "conservative")],
+              group = dados$job)
>
> # Funções de classificação
> fit$functions
             customer service    dispatch    mechanic
constant          -23.0975206 -20.7714819 -25.2062526
outdoor             0.6262915   0.8427028   0.9965617
social              1.2406066   0.7221426   1.0379702
conservative        0.6864765   1.1056750   0.7955794
> 
```

- Serviço ao Cliente: -23,10 + 1,24*22 + 0,63*10 + 0,69*5 = $\boxed{13{,}89}$
- Mecânica: -25,21 + 1,04*22 + 1,00*10 + 0,80*5 = 11,57
- Despachante: -20,77 + 0,72*22 + 0,84*10 + 1,11*5 = 9,06

O funcionário seria classificado no grupo de Serviço ao Cliente.

Classificação

A **Matriz de classificação** indica nas linhas as classificações observadas e nas colunas as classificações preditas pelo modelo. Dessa forma, as classificações corretas se encontram na diagonal da matriz.

Exemplo:

```
Console ~/
> # Matriz de confusão
> fit$confusion
                  predicted
original            customer service dispatch mechanic
  customer service                68        4       13
  dispatch                         3       50       13
  mechanic                        16       10       67
> # Taxa de erro
> fit$error_rate
[1] 0.2418033
>
```

Aproximadamente 76% de classificações corretas.

Classificação

Para avaliar a assertividade do modelo, sugere-se a **validação cruzada**:

- Tomamos uma amostra aleatória chamada de amostra análise;

- Calculamos as funções discriminantes e funções de classificação por meio dessa amostra análise;

- Testamos e avaliamos a assertividade do modelo na amostra teste.

Exercícios

5) Voltando aos dados do exercício 1, classifique um novo objeto com $X_1 = 5\ e\ X_2 = 10$. Abaixo segue a amostra observada.

$$Grupo\ A \begin{cases} X_1: 3, 2, 4 \\ X_2: 7, 4, 7 \end{cases} \text{ e } Grupo\ B \begin{cases} X_1: 6, 2, 4 \\ X_2: 9, 7, 8 \end{cases}$$

6) Considerando os dados do exercício 2, uma nova observação com $X_1 = 1\ e\ X_2 = 3$ seria classificada em qual grupo?

$$População\ 1 \begin{cases} X_1: -2, 0, -1 \\ X_2: 5, 3, 1 \end{cases},\ População\ 2 \begin{cases} X_1: 0, 2, 1 \\ X_2: 6\ 4, 2 \end{cases}\ ePopulação\ 3 \begin{cases} X_1: 1, 0, -1 \\ X_2: -2, 0, -4 \end{cases}$$

Exercícios

7) Um programa de pós-graduação deseja alterar o método de seleção dos seus alunos para uma prova de conhecimento técnico e uma nota atribuída ao histórico escolar do candidato. Para isso, os 63 candidatos do ano anterior foram divididos em três grupos: (1) constituídos dos candidatos aprovados no programa; (2) constituído pelos candidatos não aprovados, mas que ficaram na lista de espera e (3) constituído pelos candidatos não aprovados para o programa. O objetivo do estudo é verificar se o novo método de seleção é capaz de discriminar bem os candidatos.

Os dados desse exercício estão em *dados_alunos.xlsx.*

Exercícios

- (A) Qual variável parece diferenciar mais o grupo analisando as médias?
- (B) Obtenha e interprete as funções discriminantes.
- (C) Qual o percentual da variância é explicada pelas funções discriminantes?
- (D) Avalie a significâncias das funções discriminantes.
- (E) Construa e analise o gráfico scatterplot com as funções discriminantes.
- (F) De acordo com as funções de classificação, onde você classificaria um aluno que apresentasse nota técnica igual a 15 e nota 8 no histórico escolar?
- (G) Por meio da matriz de classificação, qual a assertividade da análise?
- (H) A probabilidade de acerto ao classificar uma nova observação é igual à assertividade obtida pela matriz de classificação?

Outras classificações

- **Distância de Mahalanobis**

  - É uma medida de distância que pode ser utilizada como critério de classificação.

  - Calcula-se a distância entre a observação e o centroide de cada grupo.

  - Classifica-se a observação no grupo em que ela estiver mais próxima, ou seja, cuja distância de Mahalanobis for menor.

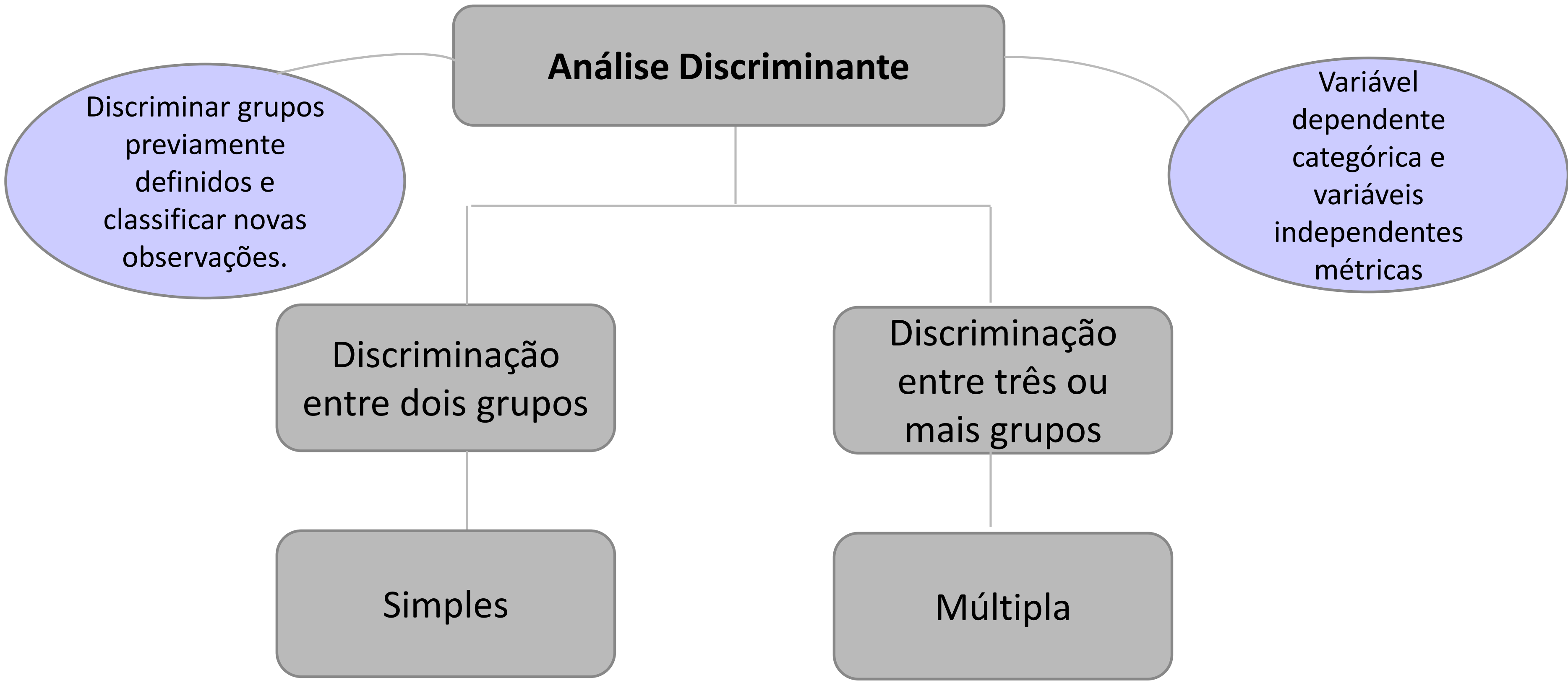
Outras classificações

- **Classificação de casos**

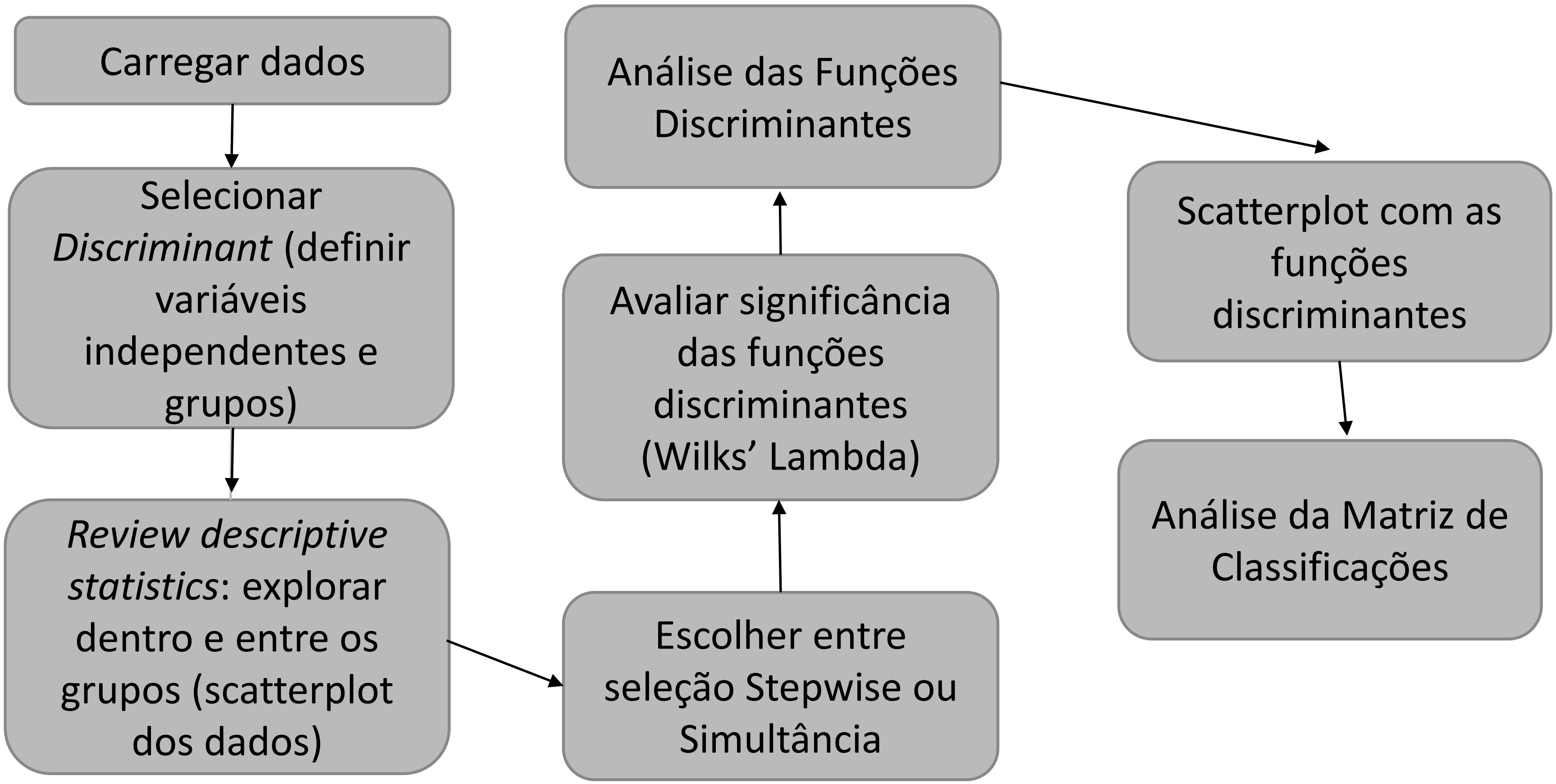
  - Podemos calcular a probabilidade de cada observação pertencer a um grupo dadas as variáveis independentes observadas. Essa é a probabilidade *a posteriori*.

  - A classificação de casos é feita classificando a observação no grupo em que ela apresenta maior probabilidade *a posteriori*.

Resumo


Análise Discriminante
Discriminar grupos previamente definidos e classificar novas observações.
Variável dependente categórica e variáveis independentes métricas
Discriminação entre dois grupos
Discriminação entre três ou mais grupos
Simples
Múltipla

Resumo


Carregar dados
Selecionar *Discriminant* (definir variáveis independentes e grupos)
Review descriptive statistics: explorar dentro e entre os grupos (scatterplot dos dados)
Escolher entre seleção Stepwise ou Simultância
Avaliar significância das funções discriminantes (Wilks' Lambda)
Análise das Funções Discriminantes
Scatterplot com as funções discriminantes
Análise da Matriz de Classificações

Referências

- BERENSON, Mark L; STEPHAN, David; LEVINE, David. Estatística: teoria e aplicações usando Microsoft excel em português. 3 ed. Rio de Janeiro: LTC – Livros Técnicos e Científicos, 2005.
- BISQUERRA, R; CASTELLA, J.; VILLEGAS, F. Introdução à estatística: enfoque informático com o pacote estatístico SPSS. Porto Alegre: Artmed, 2007.
- FAVERO, L.P.; BELFIORE, P.; SILVA, F.; CHAN, B. Análise de Dados: Modelagem Multivariada para Tomada de Decisões. Rio de Janeiro, 2010, Editora Campus.
- HAIR, J.; ANDERSON, R.; BLACK, W. Análise multivariada de dados. 5 ed. Reimp. Porto Alegre: Bookman, 2007.
- JOHNSON, R. and WICHERN, D. Applied Multivariate Statistical Analysis. Sixth edition, Wisconsin, Pearson, 2007.